HealthManagement
Promoting Management and Leadership

Volume 1 / Issue 3 2005 - Hospital IT Award

High-Quality Data Integration in Medical Information Systems

Based on the Plug-in Architecture CARAMEL

Authors

Stefan Brüggemann1 , Martin Rohde2

OFFIS – Institute for Information Technology,

Escherweg 2,26121 Oldenburg, Germany

1 brueggemann@offis.de

2 rohde@offis.de

The development and operation of the Epidemiological Cancer Registry of Lower Saxony is being supported by the institute OFFIS with the project CARLOS (CAncer Registry LOwer Saxony). In this project a couple of tools were developed and are still being maintained. CARAMEL (CARLOS Attaching Multiple Existing Local Registration Units) is used for the integration of different data sources and provides data-cleaning functionality.

CARAMEL acts as an ETL application (Extract, Transform, and Load). ETL operations are often performed in Data-Warehousing processes.

Integrating data into an information system is a common ETL task. Data has to be extracted from different data sources like databases or files. An algorithm has to be defined for each source. These algorithms must identify and read relevant data. Due to different standards and heterogenous coding systems (e.g. ICD-9, ICD-10), extracted data has to be transformed into an intermediate representation. Finally, these transformed data have to be loaded into the information system.

Challenges with Data Integration

A couple of problems had to be addressed during the integration of different data sources:

Þ Different data formats:

Medics, pathologists, and other facilities report diagnosed cancer diseases to the cancer registry. They use several different software tools to write their diagnoses. These tools use different data-export formats. Some of them even use paper-based forms, which are being sent to the cancer registry. We had to deal with a lot of changing requirements belonging to these data formats.

Þ Secure data exchange:

Data has to be transmitted to the registry only in encrypted form and using disks.

This is defined in the Law about the cancer registry. We had to guarantee the secure transmission of the diagnoses.

Þ Data Quality:

Different export formats use different codings for describing the same facts. Tools which use equal export formats interpret data fields in different ways and users might enter data into different input fields. We had to ensure the comparability of the data.

Þ Anonymity:

The cancer registry is legally obliged to preserve patient anonymity. Data stored in the information system has to be enciphered. Furthermore patient data is retrieved physically divided by the cancer registry. Patient identifying data and epidemiological data have to be separated.

Þ Flexibility:

Changing requirements for the data stored in the information system mean changes in the loading process for the database. Different studies or analyses may need different data. These data have to be collected and loaded into the database.

The Application CARAMEL

In order to fulfill the described requirements, CARAMEL was developed as a plug-in architecture. Figure 1 shows the basic structure. A plug-in architecture consists of a framework, which provides functionality for using plug-ins. Plug-ins are pieces of functionality, which can be added to or removed from the system dynamically. Plug-ins can be realised for different tasks.

Several plug-ins already are developed; others, e.g. for integrating HL7-based data (Health Level 7), will be realised when they are required.

Some data providers send data in the German BDT-format, therefore a plug-in exists for reading BDT-based files. Even dbase-database files and dc-pathos-files can be read easily. To provide a maximum of flexibility, CSV (Comma Separated Values)–based files can be read. As described above, some data providers still use paper-based forms, in which data is being entered. To deal with these data, we introduced scanner-technology. A hardware scanner is used to scan the delivered sheets and digitise them. Then an Optical Character Recognition program (OCR) extracts the included diagnoses. These texts are then being transformed into an intermediate representation.

We decided to install PGP (Pretty Good Privacy) at every data provider to ensure secure data exchange. Using this technology, medics are able to send encrypted data which can only be decrypted by the cancer registry.

Standardisation and data-cleaning are performed on the data using the internal representation. Here one has to focus on the data quality. This is a very important transformation task, because the quality of analyses carried out using the cancer registry's database directly depends on the quality of the data ("garbage in, garbage out"). Due to problems with incompatible standards or human error, data may contain dirty, missing, illegal, wrong, or abbreviated values. These syntactical errors have to be corrected automatically, whenever possible.

CARAMEL also provides interactive support for correcting files with errors manually, which cannot be repaired automatically and can provide an environment to modify incorrect values.

XSLT (eXtensible Stylesheet Language Transformations) is a flexible and extensible language for transforming data and used to support the transform- and loadtaskof the ETL-process.

Benefits of the architecture

Þ Data Provider:

Using the provided data "as is" is quite comfortable for the medics, pathologists, and others, because they do not need several tools, which results in less costs, less time for data export, and less maintenance. This is especially important with data-sources based at different locations.

Þ Plug-in Development:

Programming plug-ins is comfortable because functionality can be realised when needed and by different plug-in developers. They can be updated during runtime, for example a datacleaning component can be replaced by another one.

Þ Customisation of plug-ins:

Each data format can be used in several ways by each data source. Therefore customisable plug-ins, which can be adapted, can deal with each data source. When new data sources appear, these can easily be integrated by using an existing plug-in and plug-in configuration.

Þ Flexibility:

Using the descriptive transformation language XSLT allows for changing the transformation process on the fly. Therefore changes can be made directly at the cancer registry. This results in fast response times by developers and small maintenance costs.

Conclusion

Information systems in public health underlie several requirements on data security, anonymity and data quality. Changes in stakeholder needs often occur.

This requires a maximum of flexibility. Plug-in architectures are a convenient technology when dealing with several, frequently changing weak requirements. Changes in data formats, databases, and coding systems can be dealt with using customisable plug-ins and languages.

Paper-based forms can be dealt with, but also advanced standards like BDT or HL7.

Future Initiatives

CARAMEL has to be extended in order not only to fulfill syntactical data-cleaning tasks. Semantical and domain-specific cleaning methods have to be implemented, because they allow detection of duplicates, impossible pairs of values, dependencies, and handling of different meanings of data in specific domains. Domainspecific data-cleaning is the most advanced step in the data-cleaning process and requires metamodels, which are able to describe the Cancer Registries database and data sources.

Ontologies, which describe domain specific knowledge, may be a best-fit technique in order to model data sources and targets. Therefore work has to be done on ontology-based data cleaning.

One has to think about how to define an abstract, general architecture, which is not applied to the specific domain cancer registry, and which can be specialised in respect to concrete domains like hospitals or clinical registries.

Published on : Sat, 28 May 2005